Diffusion Maps

Máster Universitario en Ciencia de Datos - Métodos Funcionales en Aprendizaje Automático

Ángela Fernández Pascual

Escuela Politécnica Superior
Universidad Autónoma de Madrid

Academic Year 2021/22

Universidad Autónoma de Madrid

Contents

Diffusion Maps

Diffusion Maps

Method Non-linear Spectral Dimensionality Reduction method.

Goal To describe properly the geometry of the data, i.e. to find a good distance in the original sample space that characterized well the relation between the points.

New data representation that will capture the main structure of the data in a few dimensions while preserving the local structure of the original data, i.e, reducing the sample distance to a Euclidean distance.

Main assumption The manifold metric can be approximated by the **diffusion distance** of a **Markov process** whose transition matrix is defined by an adequate normalization of the similarity matrix.

Continuous-Discrete Dictionary

Continuous version		**Discrete version**	
$\mathcal{P}$	operator	$\mathbf{P}$	matrix
f	function	$\mathbf{v}$	vector
$\phi_\ell(\mathbf{x}^{(i)})$	eigenfunction	$(\phi_\ell)_i$	eigenvector
$\int$	integral	$\sum$	summation

Motivation: Diffusion Processes

Diffusion Processes (I)

Diffusion (physical context)

The process by which a gas moves from regions of high density to regions of lower density according to the relative pressure of each region.

The diffusion equation is usually written:

$$\frac{\partial \chi}{\partial t} = c\nabla^2 \chi.$$

Diffusion (graph context)

The diffusion will be a model of spread across the graph.

Examples:

- The spread of an idea in a social network.
- The spread of a disease across some region.

Diffusion Processes (II)

Assumption Undirected weighted graph, with w_{ij} the similarity between vertex i and j.

The degree of each vertex is $d_i = \sum_j w_{ij}$.

Definition Let be $\boxed{\chi_i}$ a fluid or substance located in the nodes of the graph that flows from vertex j to an adjacent vertex i with a rate $c(\chi_j - \chi_i)$.

Diffusion constant: $\boxed{c}$.

Diffusion Processes (III)- Gas Diffusion Equation

- Flow in a short period of time: $c(\chi_j - \chi_i)dt$
- χ_i changes in a ratio:

$$
\begin{aligned}
\frac{\partial \chi_i}{\partial t} &= c \sum_j w_{ij}(\chi_j - \chi_i) \\
&= c \sum_j w_{ij}\chi_j - c\chi_i \sum_j w_{ij} \\
&= c \sum_j w_{ij}\chi_j - c\chi_i d_i \\
&= c \sum_j (w_{ij} - \delta_{ij} d_i)\chi_j .
\end{aligned}
$$

- Matrix notation: $\frac{d\chi}{dt} = c(\mathbf{W} - \mathbf{D})\chi$.

Diffusion Processes (IV)- Laplacian Graph

Laplacian matrix

$\mathbf{L} = \mathbf{D} - \mathbf{W}$

We can then rewrite:

$$\boxed{\frac{\partial \chi}{\partial t} + c\mathbf{L}\chi = 0}.$$

We have arrived to the gas diffusion process equation (assuming $\mathbf{L} \approx \Delta = \nabla^2$).

Diffusion Processes (V)- Equation Solution

- Eigendecomposition of $\mathbf{L}$:
$$\mathbf{L}\phi_i = \lambda_i \phi_i.$$
- $\{\phi_i\}$ form an orthonormal basis, so:
$$\chi(t) = \sum_i w_i(t)\phi_i,$$
being $w_i(t)$ the coefficients to define χ depending on time t.
- Replacing in the diffusion equation:
$$\frac{\partial(\sum_i w_i(t)\phi_i)}{\partial t} + c\mathbf{L}\sum_i w_i(t)\phi_i = \sum_i \left(\frac{\partial w_i(t)}{\partial t} + c\lambda_i w_i(t)\right)\phi_i = 0.$$

Diffusion Processes (VI)- Equation Solution

Rewritten equation $\frac{\partial w_i(t)}{dt} + c\lambda_i w_i(t) = 0$

Solution $w_i(t) = w_i(0)e^{c\lambda_i t}$.

The flux can be determined under some initial conditions $w_i(0)$, just computing the eigendecomposition of $\mathbf{L}$.

Diffusion Maps is a particular type of diffusion process that can be used to study the underlying relationship between points in a data set.

Diffusion Maps Algorithm

Defining Diffusion Coordinates: Building a Graph

- Starting sample: $\mathscr{S} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$.
- A diffusion process over the data $\Rightarrow$ **Random Walk** over a graph.
- First step: to build a **symmetric weighted graph**.

Affinity matrix

In terms of a kernel matrix $\mathbf{K}$ which results from a kernel operator $\mathcal{K} : \mathscr{S} \times \mathscr{S} \to \mathbb{R}$, defining the matrix entries as $k_{ij} = \mathcal{K}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$.

Properties:

- Symmetric: $k_{ij} = k_{ji}$.
- Positive: $k_{ij} \geq 0$.
- Positive semi-definite: $\mathbf{x}^{(i)\top} \mathbf{K} \mathbf{x}^{(i)} \geq 0 \quad \forall \mathbf{x}^{(i)} \in \mathscr{S}$.

- How to chose the kernel? It depends on the concrete problem.

  $\Rightarrow$ Usual choice: **Gaussian kernel** defined as $k_{ij} = e^{\frac{-\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2}{2\sigma^2}}$.

Notebook

Building a Graph

Random Walk over the graph (I)

- We need to build now a **Markov Chain** (MC) over the graph.
- In particular, we build a **random walk**.
  - It defines a local relationship between points in the graph providing a structure to the data.

Normalized Graph Laplacian construction

We define a **transition probability** from the affinity matrix.

1. Degree of each node in the graph as:

$$d_i = \sum_{j \in \mathscr{S}} k_{ij}.$$

2. Transition probability $\equiv$ normalized kernel:

$$p_{ij} = \frac{k_{ij}}{d_i}.$$

Random Walk over the graph (II)

Is this transition probability of the random walk well-defined?

1. $\mathbf{P} \geq 0$
2. $\sum_j p_{ij} = \frac{\sum_j k_{ij}}{d_i} = \frac{d_i}{d_i} = 1.$

- $\mathbf{P} \equiv$ probability of arriving from i to j in one step.
- $\mathbf{P}^{\tau} = (p_{ij}^{\tau}) \equiv$ probability of arriving from i to j in τ steps.
- $\boxed{\tau}$: scale.
    - Running the process far away in time (with a higher τ) lets us integrate the local geometry, so the structure of the data is revealed at different scales.

Notebook

Random Walk over the graph: Transition Probability and steps

Random Walk over the graph (III)

Stationary Distribution of the MC: $\boxed{\pi_i = \frac{d_i}{\sum_k d_k}}$

Properties of π_i

1. We can prove that it is **stationary**:

$$\sum_i \pi_i p_{ij} = \sum_i \frac{d_i}{\sum_k d_k} \frac{k_{ij}}{d_i} = \sum_i \frac{k_{ij}}{\sum_k d_k} = \frac{d_j}{\sum_k d_k} = \pi_j.$$

2. It is a **reversible** chain, as it satisfies the balance equations:

$$\pi_i p_{ij} = \frac{d_i}{\sum d_k} \frac{k_{ij}}{d_i} = \frac{k_{ij}}{\sum d_k} = \frac{k_{ji}}{\sum d_k} = \frac{d_j}{\sum d_k} \frac{k_{ji}}{d_j} = \pi_j p_{ji}.$$

3. The chain is **irreducible**, as the graph is connected.
4. The chain is **aperiodic**: $g(\mathbf{x}^{(i)}) = \gcd\{\tau \geq 1 : p_{ii}^{\tau} > 0\} = 1 \ \forall i$.

$\Rightarrow$ The MC es **ergodic**.

Notebook

Random Walk over the graph: MC Distribution π

Diffusion Distance (I)

Final objective To find a good distance in the original sample space that is reduced to a Euclidean distance.

Possible distance Based on the previous Markov chain.

$$
\begin{aligned}
\boxed{\mathcal{D}_{\tau}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})} &= \| p^{\tau}_{i\cdot} - p^{\tau}_{j\cdot} \|^2_{(1/\pi)} = \sum_k \frac{\left(p^{\tau}_{ik} - p^{\tau}_{jk}\right)^2}{\pi_k} \\
&= \sum_k \frac{(p^{\tau}_{ik})^2 + (p^{\tau}_{jk})^2 - 2p^{\tau}_{ik}p^{\tau}_{jk}}{\pi_k} = \sum_k \frac{p^{\tau}_{ik}p^{\tau}_{ik} + p^{\tau}_{jk}p^{\tau}_{jk} - p^{\tau}_{ik}p^{\tau}_{jk} - p^{\tau}_{jk}p^{\tau}_{ik}}{\pi_k} \\
&= \sum_k \frac{1}{\pi_k}\left(p^{\tau}_{ik}p^{\tau}_{ki}\frac{\pi_k}{\pi_i} + p^{\tau}_{jk}p^{\tau}_{kj}\frac{\pi_k}{\pi_j} - p^{\tau}_{ik}p^{\tau}_{kj}\frac{\pi_k}{\pi_j} - p^{\tau}_{jk}p^{\tau}_{ki}\frac{\pi_k}{\pi_i}\right) \\
&= \frac{1}{\pi_i}\sum_k p^{\tau}_{ik}p^{\tau}_{ki} + \frac{1}{\pi_j}\sum_k p^{\tau}_{jk}p^{\tau}_{kj} - \frac{1}{\pi_j}\sum_k p^{\tau}_{ik}p^{\tau}_{kj} - \frac{1}{\pi_i}\sum_k p^{\tau}_{jk}p^{\tau}_{ki} \\
&= \frac{p^{2\tau}_{ii} - p^{2\tau}_{ji}}{\pi_i} + \frac{p^{2\tau}_{jj} - p^{2\tau}_{ij}}{\pi_j}.
\end{aligned}
$$

Diffusion Distance (II)

We will call this distance **Diffusion Distance**.

Properties of the Diffusion Distance

1. $\mathcal{D}_\tau(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ is symmetric.
2. $\mathcal{D}_\tau(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ satisfies the triangle inequality.

$\Rightarrow$ It is a semi-metric.

$\Rightarrow$ If $k_{ij} > 0 \; \forall i, j$, it is a metric.

3. It is robust to noise (calculated as an average of all the paths of length τ).

The **diffusion distance** measures the connectivity between two points in the dataset after 2τ steps.

Intuition: $\mathcal{D}_\tau(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ small if there exist a lot of paths that connect $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$

$\Rightarrow$ when $p_{ij}^{2\tau}$ and $p_{ji}^{2\tau}$ are high.

Spectral Theory (I)

Theorem (Spectral Theorem)

Any symmetric matrix $\mathbf{A} \in \mathbb{R}^n \times \mathbb{R}^n$ *can be reduced to an orthonormal basis determined by a diagonal matrix, i.e.,* $\mathbf{A} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}'$, *where*

- $\mathbf{\Lambda} = diag\{\lambda_0, \lambda_1, \cdots, \lambda_N\}$, *with* λ_ℓ *the eigenvalues of* $\mathbf{A}$,
- $\mathbf{\Phi} = (\phi_0, \cdots, \phi_N)$, *with* ϕ_ℓ *the eigenvectors of* $\mathbf{A}$.

! Our Markov matrix **P** is not symmetric.

→ Let's consider its conjugated matrix **A**:

$$a_{ij} = \frac{\sqrt{\pi_i}}{\sqrt{\pi_j}} p_{ij} = \frac{\sqrt{\pi_i}}{\sqrt{\pi_j}} \frac{k_{ij}}{d_i} = \frac{\sqrt{\frac{d_i}{\sum d_k}}}{\sqrt{\frac{d_j}{\sum d_k}}} \frac{k_{ij}}{d_i} = \frac{k_{ij}}{\sqrt{d_i}\sqrt{d_j}}$$

Spectral Theory (II)

- Applying the Spectral Theorem over $\mathbf{A}$: $\boxed{a_{ij} = \sum_{\ell \geq 0} \lambda_\ell (\phi_\ell)_i (\phi_\ell)_j}$
  - $\{\lambda_\ell\}_{\ell \geq 0}$, eigenvalues of $\mathbf{A}$
  - $\{\phi_\ell\}_{\ell \geq 0}$, eigenvector of $\mathbf{A}$.
- $\boxed{\lambda_0 = 1}$ corresponds to the eigenvector $\phi_0 = \sqrt{\pi}$.

$$(\mathbf{A}\sqrt{\pi})_i = \sum_j a_{ij}\sqrt{\pi_j} = \sum_j \frac{\sqrt{\pi_i}}{\sqrt{\pi_j}} p_{ij}\sqrt{\pi_j} = \sqrt{\pi_i}\sum_j p_{ij} = \sqrt{\pi_i}\mathbf{1} = \sqrt{\pi_i} \quad \forall i.$$

Spectral Theory (III)

- $\mathbf{P}$ rewritten:

$$p_{ij} = \frac{\sqrt{\pi_j}}{\sqrt{\pi_i}} a(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \frac{\sqrt{\pi_j}}{\sqrt{\pi_i}} \sum_{\ell \geq 0} \lambda_\ell (\phi_\ell)_i (\phi_\ell)_j = \sum_{\ell \geq 0} \lambda_\ell \frac{(\phi_\ell)_i}{\sqrt{\pi_i}} (\phi_\ell)_j \sqrt{\pi_j}.$$

$\Rightarrow$ Eigendecomposition of $\mathbf{P}$ for any τ-steps:

$$\boxed{p_{ij} = \sum_{\ell \geq 0} \lambda_\ell^\tau (\psi_\ell)_i (\varphi_\ell)_j,}$$

where $(\psi_\ell)_i = \frac{(\phi_\ell)_i}{\sqrt{\pi_i}}$ and $(\varphi_\ell)_j = (\phi_\ell)_j \sqrt{\pi_j}$.

- $\lambda_0 = 1$ discarded also for $\mathbf{P}$.
- The other eigenvalues will satisfy $|\lambda_\ell| < 1 \ \forall \ell \geq 1$.

Simplifying $\mathcal{D}_T(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$

$$\boxed{\mathcal{D}_T(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})} = \|p_{ik}^T - p_{jk}^T\|^2_{(1/\pi)} = \sum_k \frac{\left(p_{ik}^T - p_{jk}^T\right)^2}{\pi_k}$$

$$= \sum_k \frac{\left(\sum_{\ell\geq 0} \lambda_\ell^T (\psi_\ell)_i (\varphi_\ell)_k - \sum_{\ell\geq 0} \lambda_\ell^T (\psi_\ell)_j (\varphi_\ell)_k\right)^2}{\pi_k}$$

$$= \sum_k \frac{\sum_{\ell\geq 0} \lambda_\ell^{2T} \left((\psi_\ell)_i - (\psi_\ell)_j\right)^2 (\varphi_\ell)_k^2}{\pi_k} = \sum_{\ell\geq 0} \lambda_\ell^{2T} \left((\psi_\ell)_i - (\psi_\ell)_j\right)^2 \sum_k \frac{(\varphi_\ell)_k^2}{\pi_k}$$

$$= \sum_{\ell\geq 0} \lambda_\ell^{2T} \left((\psi_\ell)_i - (\psi_\ell)_j\right)^2 \sum_k \frac{\left((\phi_\ell)_k \sqrt{\pi_k}\right)^2}{\pi_k}$$

$$= \boxed{\sum_{\ell\geq 0} \lambda_\ell^{2T} \left((\psi_\ell)_i - (\psi_\ell)_j\right)^2}.$$

Diffusion Coordinates

- **Dimensionality Reduction**: $\lambda_\ell \to 0$ when ℓ grows (small contribution to the Diffusion Distance).
  - Given a precision δ, we work with:

$$\bar{M} = s(\delta, \tau) = \max\{\ell \in \mathbb{N} \text{ s.t. } |\lambda_\ell|^\tau > \delta |\lambda_1|^\tau\}.$$

  - $\ell = 0$ is omitted as ψ_0 is constant.
- **Diffusion distance approximation**:

$$\mathcal{D}_\tau(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \approx \sum_{\ell \geq 1}^{\bar{M}} \lambda_\ell^{2\tau} \left((\psi_\ell)_i - (\psi_\ell)_j\right)^2 .$$

Diffusion Coordinates

They are the natural coordinates for representing the original sample, and they are defined as:

$$\mathbf{\Psi}_\tau(\mathbf{x}) = \begin{pmatrix} \lambda_1^\tau \psi_1(\mathbf{x}) \\ \vdots \\ \lambda_{\bar{M}}^\tau \psi_{\bar{M}}(\mathbf{x}) \end{pmatrix} .$$

Diffusion Maps

Diffusion Maps

We call **Diffusion Map** to the family $\{\mathbf{\Psi}_T\}_{T\in\mathbb{N}}$.

- These maps embed the data into the Euclidean space $\mathbb{R}^{\bar{M}}$:

$$
\begin{aligned}
\|\mathbf{\Psi}_T(\mathbf{x}^{(i)}) - \mathbf{\Psi}_T(\mathbf{x}^{(j)})\| &= \sum_{\ell\geq 1}^{\bar{M}} \lambda_\ell^{2T} \left((\psi_\ell)_i - (\psi_\ell)_j\right)^2 \\
&\approx \mathcal{D}_T(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \sum_{\ell\geq 1}^{\bar{M}} \lambda_\ell^{2T} \left((\psi_\ell)_i - (\psi_\ell)_j\right)^2 + \sum_{\ell>\bar{M}}^{N} \lambda_\ell^{2T} \left((\psi_\ell)_i - (\psi_\ell)_j\right)^2 .
\end{aligned}
$$

Diffusion Maps Algorithm Summary

Original Data $\mathscr{S}$

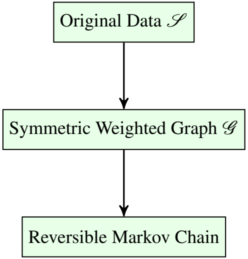
Diffusion Maps Algorithm Summary

$$\begin{array}{c}\boxed{\text{Original Data } \mathscr{S}}\\ \downarrow \\ \boxed{\text{Symmetric Weighted Graph } \mathscr{G}}\end{array}$$

- Kernel definition:

$$\mathbf{K} = \{k_{ij}\} = \left\{ e^{\frac{-\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2}{2\sigma^2}} \right\}.$$

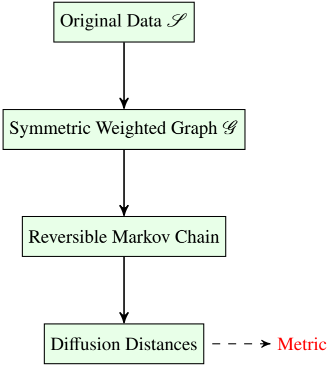
Diffusion Maps Algorithm Summary



- Transition Probability Matrix definition:

$$p_{ij} = \left(\frac{k_{ij}}{d_i} \right) = \left(\frac{k_{ij}}{\sum_j k_{ij}} \right).$$

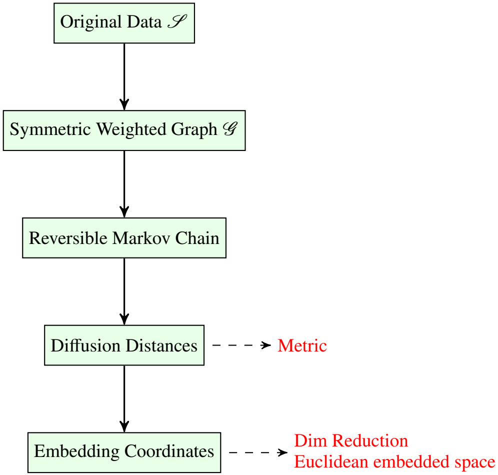
Diffusion Maps Algorithm Summary



- Diffusion Distance:

$$\mathcal{D}_T(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \|p_{i\cdot}^T - p_{j\cdot}^T\|^2_{(1/\pi)}$$
$$= \frac{p_{ii}^{2T} - p_{ji}^{2T}}{\pi_i} + \frac{p_{jj}^{2T} - p_{ij}^{2T}}{\pi_j}.$$

Diffusion Maps Algorithm Summary



- Spectral Theory:

$$p_{ij}^{T} = \sum_{\ell \geq 0} \lambda_\ell^{T} (\psi_\ell)_i (\varphi_\ell)_j$$

$$\rightarrow \mathcal{D}_T(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \sum_{\ell=0}^{M} \lambda_\ell^{2T} \left((\psi_\ell)_i - (\psi_\ell)_j\right)^2 .$$

- Dimensionality Reduction:

$$\bar{M} = \max\{\ell \in \mathbb{N} \text{ s.t. } |\lambda_\ell|^{T} > \delta |\lambda_1|^{T}\}.$$

Diffusion Maps Algorithm Summary

Original Data $\mathscr{S}$

↓

Symmetric Weighted Graph $\mathscr{G}$

↓

Reversible Markov Chain

↓

Diffusion Distances ⇢ Metric

↓

Embedding Coordinates ⇢ Dim Reduction
Euclidean embedded space

- Diffusion Coordinates Definition:

$$\Psi_{\mathrm{T}}(\mathbf{x}) = \begin{pmatrix} \lambda_1^{\mathrm{T}} \psi_1(\mathbf{x}) \\ \vdots \\ \lambda_{\bar{M}}^{\mathrm{T}} \psi_{\bar{M}}(\mathbf{x}) \end{pmatrix}.$$

- Euclidean Distances over the embedding space:

$$\begin{aligned} \left\|\Psi_{\mathrm{T}}(\mathbf{x}^{(i)}) - \Psi_{\mathrm{T}}(\mathbf{x}^{(j)})\right\| &= \sum_{\ell \geq 1}^{\bar{M}} \lambda_\ell^{2\mathrm{T}} \left((\psi_\ell)_i - (\psi_\ell)_j\right)^2 \\ &\approx \mathcal{D}_{\mathrm{T}}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}). \end{aligned}$$

Density Influence

Density Influence

- The distribution of the sample data affects how the similarity matrix captures the local geometry of the data.
- The density is measured by the degree of the graph **D**.
- We introduce a new parameter $\boxed{\alpha \in [0, 1]}$.
- Steps for normalizing:
  1. Define α-dependent density normalized matrix:
  $$k_{ij}^{(\alpha)} = \frac{k_{ij}}{d_i^\alpha d_j^\alpha}.$$
  2. Classical normalization:
  $$p_{ij}^{(\alpha)} = \frac{k_{ij}^{(\alpha)}}{d_i^{(\alpha)}}.$$

Density Influence - Importance of α

$$\nabla_\alpha f = \frac{\Delta(f\mathbf{d}^{1-\alpha})}{\mathbf{d}^{1-\alpha}} - \frac{\Delta(\mathbf{d}^{1-\alpha})}{\mathbf{d}^{1-\alpha}} f$$

$\alpha = 1$ $\nabla_1 \equiv \Delta$

$\rightarrow$ DM captures the underlying geometry without interference from the sample's density $\mathbf{d}$.

$\alpha = 0$ $\nabla_0 f = \frac{\Delta(f\mathbf{d})}{\mathbf{d}} - \frac{\Delta(\mathbf{d})}{\mathbf{d}} f$

$\rightarrow$ $\mathbf{d}$ influences how the diffusion coordinates capture the underlying geometry, unless $\mathbf{d}$ is uniform ($\nabla_0 = \Delta$).

Tips on Practical Use

Tips on Practical Use (I)

1. The scale must be the same over all the features $\rightarrow$ Scale the data ($\mu = 0, \sigma = 1$).
2. Good hyper-parameter selection.

Affinity Matrix

- Gaussian kernel:

$$k_{ij} = e^{\frac{-\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2}{2\sigma^2}}.$$

- σ is crucial, it can be chose as a distance percentile (ρ):

$$\sigma = p_\rho\left(\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|\right)$$

  - Usual ρ:
    - Median.
    - Maximum.

Tips on Practical Use (II)

Reduced dimension

For a given step τ:

$$\bar{M} = s(\delta, \tau) = \max\{\ell \in \mathbb{N} \text{ s.t. } |\lambda_\ell|^\tau > \delta|\lambda_1|^\tau\}.$$

Number of steps

By default, $\tau = 1$.

Density influence

By default, $\alpha = 1$.

Out-of-sample Extension

DM Advantages and Disadvantages

Advantages

- Powerful
- Elegant

Disadvantages

- Costly eigenanalysis of the transition matrix.
- The extension to new, unseen points (out-of-sample points) is not possible.

The Nyström Formula

- **Nyström**: approximate the eigenfunctions $\phi_j(\mathbf{x}^{(i)})$ of a symmetric and positive semidefinite kernel from the eigenvectors $(\phi_j)_i$ of a sample-based kernel matrix.
  $\Rightarrow$ It maintains the eigenvalue and eigenvector convergence as the number of sample patterns grows.

- The Nyström Formula:

$$\boxed{v_j(\mathbf{y}) = \frac{1}{\lambda_j} \sum_{i=1}^{n} v_j(\mathbf{x}^{(i)}) \mathcal{K}(\mathbf{y}, \mathbf{x}^{(i)})}$$

The Nyström Formula applied to DM (I)

- Nyström approximates the DM embedding of new patterns
  - $\Rightarrow$ No need to repeat the eigendecomposition of $\mathbf{P}$ over the training sample.
- Problem: $\mathbf{P}$ is not symmetric.
  - $\Rightarrow$ Let's use its conjugate matrix $\mathbf{A}$.
- λ_j and $\phi_j(\mathbf{x}^{(i)})$: eigenvalues and eigenfunctions of $\mathbf{A}$ of the sample data.
- Eigenvector extension for a new point $\mathbf{x}$:

$$\phi_j(\mathbf{x}) = \frac{1}{\lambda_j} \sum_{i=1}^{N} \phi_j(\mathbf{x}^{(i)}) a(\mathbf{x}, \mathbf{x}^{(i)}).$$

The Nyström Formula applied to DM (II)

Recall

- $a(\mathbf{x}, \mathbf{x}^{(i)}) = \frac{\pi(\mathbf{x})}{\pi(\mathbf{y})} \mathbf{P}(\mathbf{x}, \mathbf{x}^{(i)})$.
  - Eigenvalues: (λ_ℓ).
  - Eigenvectors: (ϕ_ℓ).
- **P** left eigenvector: $(\psi_\ell)_i = \frac{(\phi_\ell)_i}{\sqrt{\pi_i}}$
- **P** right eigenvector: $(\varphi_\ell)_j = (\phi_\ell)_j \sqrt{\pi_j}$

The Nyström Formula applied to DM (III)

- Nyström formula for a new $\mathbf{x}$ yields:

$$\psi_j(\mathbf{x}) = \frac{\phi_j(\mathbf{x})}{\sqrt{\pi(\mathbf{x})}} = \frac{1}{\sqrt{\pi(\mathbf{x})}} \left(\frac{1}{\lambda_j} \sum_{i=1}^{N} \phi_j(\mathbf{x}^{(i)}) a(\mathbf{x}, \mathbf{x}^{(i)}) \right)$$
$$= \frac{1}{\lambda_j} \sum_{i=1}^{N} \phi_j(\mathbf{x}^{(i)}) \frac{a(\mathbf{x}, \mathbf{x}^{(i)})}{\sqrt{\pi(\mathbf{x})}} = \frac{1}{\lambda_j} \sum_{i=1}^{N} \phi_j(\mathbf{x}^{(i)}) \frac{\mathbf{P}(\mathbf{x}, \mathbf{x}^{(i)})}{\sqrt{\pi\left(\mathbf{x}^{(i)}\right)}}$$
$$= \frac{1}{\lambda_j} \sum_{i=1}^{N} \psi_j(\mathbf{x}^{(i)}) \mathbf{P}(\mathbf{x}, \mathbf{x}^{(i)}).$$

$\Rightarrow$ The Nyström formula can be applied almost in its original form to $\mathbf{P}$.

- Moreover, this formulation is also valid for any α parameter value used to compute any of our $\mathbf{P}^{(\alpha)}$ matrices to be used in DM.

References

References

- Coifman, R. R., Lafon, S. (2006). **Diffusion maps**. Applied and computational harmonic analysis, 21(1), 5-30
- Nadler, B., Lafon, S., Coifman, R. R., Kevrekidis, I. G. (2006). **Diffusion maps, spectral clustering and reaction coordinates of dynamical systems**. Applied and Computational Harmonic Analysis, 21(1), 113-127.
- Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., Zucker, S. W. (2005). **Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps**. Proceedings of the national academy of sciences, 102(21), 7426-7431.